UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

March 22, 2001

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES # Q-43

MEMORANDUM FOR The Record

From: Howard Hogan
Chief, Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation Survey:
Effect of Excluding "Late Census Adds"

I. Introduction

During the census process, 2.3 million person records in the U.S. were excluded from the Accuracy and Coverage Evaluation (A.C.E.) processing and estimation, but later included in the Hundred-percent Census Edited File (HCEF). This memorandum discusses the effects of the decision not to include these cases in A.C.E. processing and estimation.

The discussion below will first focus on the effect of these removed cases on the A.C.E. expected value, in order to discuss any possible effect on bias. The next section will give a simple example. A final section will discuss the effect on variance. The intent of this discussion is to explain these effects in layman's terms. Consequently, rigorous statistical derivation and terminology are purposefully avoided.

II. Effect on Bias

This section builds a model which is then used to answer the question: What is the effect of not including these cases in the Dual System Estimator (DSE) used to estimate undercount and adjustment? The DSE can be expressed as a function of three factors:

- The number of complete and correct census enumerations in the census file used for A.C.E. processing and estimation (N_{+1}).
- The number of persons in the A.C.E. (N_{1+}).

- The number of persons in common (N_{11}) among the N_{+1} and the N_{1+}. (M is used to denote the number of complete and correct census persons who are also in the A.C.E.)

Specifically, the DSE can be written as:

$$DSE = \frac{N_{1+} N_{+1}}{N_{11}} = N_{1+} \frac{N_{+1}}{N_{11}} . \qquad (1)$$

In other words, the DSE can be written as either:

- The product of the number of people in each of the two systems divided by the number of people in both, or as

- The product of the number of A.C.E. people and the ratio of census complete and correct enumerations to the number of people in both systems.

Obviously, any operation that does not affect either the number of A.C.E. people (N_{1+}) or the ratio of census complete and correct enumerations to the number of people in both systems (N_{+1} / N_{11}) will not affect the DSE. Conversely, any effect must come through one of these two factors.

Having estimated the true population via the DSE, the A.C.E. results can be used to compute two measures of net census error. The first is the Net Percent Undercount (UCR), written as

$$UCR\,\% = 100 \times \frac{DSE - C}{DSE},$$

while the second measure is the Coverage Correction Factor (CCF), displayed as

$$CCF = \frac{DSE}{C}.$$

In both equations, C represents the final census count. The CCF is used for synthetic estimation. Again, unless an action affects either the DSE or the census totals, it cannot affect either the net percent undercount or the coverage correction factor.

Following is a brief description of the terms included in the DSE. The A.C.E. operationalizes the DSE through the estimators described below. The number of people correctly in the census files, N_{+1} , is estimated by the census count minus the number of census whole person imputations, less

the number of late census adds (i.e. cases not available for A.C.E. processing), less the number of census erroneous enumerations, shown as

$$N_{+1} = C - II - LA - EE, \qquad (2)$$

where

C = Final census count
II = Whole person imputations
LA = Late Adds (cases excluded from A.C.E. processing)
EE = Incomplete or erroneous (data-defined) census records.

Stated another way, if we define the number of data-defined census records as

$$DD = C - II - LA,$$

then

$$N_{+1} = C - II - LA - EE = DD - EE = CE. \qquad (3)$$

The number of census erroneous enumerations (EE) is usually estimated by:

$$EE = DD - DD\left(\frac{CE}{NE}\right)$$

where

NE = weighted number of E sample persons, and
CE = weighted number of correct enumerations.

However, in expectation, the weighted number of E sample persons is equal to the number of data-defined census records in the A.C.E. universe, i.e. E (NE) = DD. As a result,

$$EE = DD - DD\left(\frac{CE}{DD}\right) = DD - CE.$$

We now focus on the effect of the excluded records (Late Adds) on the dual system estimate. Specifically, what would their effect have been if they were available for A.C.E. matching and processing?

If the LA cases had been processed, they would have fallen into one of several mutually exclusive categories:

$$II' = \text{Whole person imputations}$$
$$EE' = \text{Incomplete or erroneous enumerations}$$
$$CE' = \text{Correct enumerations}.$$

Note that $LA = II' + EE' + CE'$. From (3), it should be clear that moving cases from LA to either II or EE cannot affect the estimate of true population via the DSE. (Recall that our discussion is limited to DSE expected values or effects on bias.) Therefore, we only need to discuss how processing correct enumerations among the LAs might affect the DSE.

Denote the number of additional correct enumerations among the Late Adds that would have matched to the P sample had they been included as

$$M' = \text{Matched to the P sample}$$

Using this notation, we compute

$$N'_{+1} = C - II - II' - EE - EE' = CE + CE',$$

$$N'_{11} = M + M',$$

and

$$DSE' = N_{1+} \frac{N'_{+1}}{N'_{11}}. \qquad (4)$$

It should be clear that excluding Late Adds from A.C.E. processing will not affect the number of A.C.E. people (N_{1+}). It follows that the two estimators will be equal if and only if:

$$DSE' = DSE$$

$$\Leftrightarrow \quad N_{1+} \frac{N'_{+1}}{N'_{11}} = N_{1+} \frac{N_{+1}}{N_{11}}$$

$$\Leftrightarrow \quad \frac{N'_{+1}}{N'_{11}} = \frac{N_{+1}}{N_{11}}$$

$$\Leftrightarrow \qquad \frac{CE + CE'}{M + M'} = \frac{CE}{M}$$

$$\Leftrightarrow \qquad \frac{CE'}{M'} = \frac{CE}{M}.$$

In other words, excluding the Late Adds will not affect the DSE of the true population if the number of matches is reduced proportionately to the number of census correct enumerations. Said another way, the probability of a Late Add being excluded from A.C.E. processing must be statistically independent of its inclusion probability in the A.C.E. This is, of course, the traditional dual system independence assumption.

III. An Example

A simple example may help. See the attached table to follow this example. Consider first the estimator if there were no Late Adds. That is, suppose all cases had been available for A.C.E. processing and estimation.

Assume that in a post-stratum, the census count is one million, of which 30,000 are whole person imputations. Since we are dealing with expected values, the E-sample total is equal to the census number of data defined records, i.e. the E-sample frame. Let us assume for this example that there are 48,500 erroneous enumerations, yielding 921,500 correct enumerations.

For our example, the P sample total is 950,000. Of these, 95,000 do not match to the census while 855,000 do. This gives a census coverage ratio (census gross completeness as measured by the A.C.E.) of 90 percent. It also gives an A.C.E. coverage ratio (A.C.E. gross completeness as measured by the census) of 92.78 percent.

The usual DSE is calculated as CE (N_{1+} / M), or in this example 1,023,889. This number is compared to the census total of one million for a net undercount of 2.33 percent and a coverage correction factor (CCF) of 1.0239.

Now, what is the effect of excluding some of the census cases from A.C.E. processing? In this example, one percent (10,000) of the records are excluded. Some of these records (1,000) are whole person imputations while others are duplicate or other erroneous records (4,000). Thus, the number of census correct enumerations is reduced by only 5,000, or about one half of one percent.

Removing correct enumerations will reduce the number of matches (4,639) and increase the number of A.C.E. non-matches by the same (4,639) amount. Note that this example is constructed such that the A.C.E. coverage ratio remains fixed at 92.78 percent.

Under this assumption, the DSE remains exactly the same.[1] So if the estimated true population remains unchanged and the census total remains unchanged, both the net undercount rate and the coverage correction factor remain unchanged.

In order to introduce an appreciable bias, **two** conditions must occur:

1. A large proportion of correct enumerations must have been excluded, and

2. These correctly enumerated people must have a different probability of inclusion in the A.C.E. than the non-excluded cases.

For example, suppose two percent (20,000) of the census cases were Late Adds. Further, assume that half of them (10,000) were either whole person imputations or otherwise erroneous enumerations.

- If **all** 10,000 of these cases had matched to the A.C.E., there would have been 10,000 fewer matches , resulting in 845,000 matches (M). Excluding these cases from the DSE has the effect of increasing the estimated net percent undercount from 2.33 to 2.42.

- Further, if **only half** (5,000) of these cases had matched to the A.C.E., removing them would have lowered the estimated net percent undercount from 2.33 to 1.84.

These are extreme assumptions. Certainly, the Late Adds were set aside because they were more likely to be duplicates, i.e. to be erroneous enumerations. Further, it is hard, on the one hand, to believe that the A.C.E. could possibly include all excluded correct enumerations. It is equally hard, on the other hand, to believe that these individuals were significantly less likely to have been found by the A.C.E. than other people correctly included in the census. No causal mechanism has yet to be suggested. Because of this, we see no reason to believe that excluding these cases from A.C.E. processing introduced a significant bias.

[1] Actually, if the calculation is carried out using integers, there will be a very small difference due to rounding.

IV. Effect on Variance

The above discussion has focused solely on the expected value or the bias. We now address the issue of variance. Excluding Late Adds from the A.C.E. has opposing effects on the variance; one decreases it while another increases it.

In order to produce the DSE, we must estimate the number of EE's from the E sample. Removing EE's on a 100 percent (non-sample) basis should decrease the variance, especially if the removed EE's are clustered. Thus, to the extent that the excluded cases included large numbers of clustered erroneous enumerations, e.g. person duplicates, the variance of the DSE will be reduced.

However, some correct enumerations were also removed and not included in the A.C.E. processing. Again, if these CEs are clustered, they would lead to clusters of A.C.E. non-matches. This would increase the variance of the DSE.

Fortunately, both of these effects should be included in the DSE variance estimates that will be produced. It will not be possible to disaggregate the variance and attribute portions of the variance to any particular cause, including the Late Adds. However, the net effect of these processes on the variance will be properly accounted for.

Table 1. Example of Effect of Late Adds on DSE

Expected Estimates		DSE includes Late Adds	Late Adds	DSE excludes Late Adds
Census Count	C	1,000,000		1,000,000
less				
Late Census Additions	LA	0	10,000	10,000
Whole Person Imputations	II	30,000	-1,000	29,000
equals				
Data Defined	**DD**	**970,000**		**961,000**
E-Sample (Expected Value)	NE	970,000		961,000
less				
Erroneous Enumerations	EE	48,500	-4,000	44,500
equals				
Correct Enumerations	**CE**	**921,500**	**-5,000**	**916,500**
P-Sample Total	N_{1+}	950,000		950,000
less				
Non-matches	$N_{1+} - M$	95,000	4,639	99,639
equals				
Matches	M	855,000	-4,639	850,361
Census Coverage Ratio	**M/N_{1+}**	**0.9000**		**0.8951**
A.C.E. Coverage Ratio	M /CE	0.9278		0.9278
Estimated Population **DSE**	**$CE(N_{1+}/M)$**	**1,023,889**		**1,023,889**
Census Total	C	1,000,000		1,000,000
Coverage Correction Factor	**CCF**	**1.0239**		**1.0239**
Net Undercount	DSE - C	23,889		23,889
Net Percent Undercount	(DSE - C)/DSE	2.33		2.33